Roles of the Average Voice in Speaker-adaptive HMM-based Speech Synthesis

Junichi Yamagishi[1], *Oliver Watts*[1], *Simon King*[1], *Bela Usabaev*[2]

[1]The Centre for Speech Technology Research,
University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom
[2]Universität Tübingen, Wilhelmstr. 7 72074 Tübingen, Germany

jyamagis@inf.ed.ac.uk

Abstract

In speaker-adaptive HMM-based speech synthesis, there are a few speakers whose synthetic speech sounds worse than that of other speakers, despite having the same amount of adaptation data from within the same corpus. This paper investigates these fluctuations in quality and found that as mel-cepstral distance from the average voice becomes larger, the MOS scores generally become worse. Although the negative correlation obtained is not strong enough, this helps us improve the training and adaptation strategies for average voice models. Furthermore we remark that this correlation is strongly linked to "vocal attractiveness."

Index Terms: speech synthesis, HMM, average voice, speaker adaptation

1. Introduction

Until recently, developing a text-to-speech synthesis system for a targeted speaker required a large amount of speech data from a carefully prepared script. However, with the advent of the HMM-based speech synthesis system [1], statistical acoustic models for spectral, excitation, and duration features can now be precisely adapted from an average voice model (derived from other speakers) or a background model (derived from one speaker) using only a very small amount of speech data.

Recent experiments with the speaker-adaptive HMM-based speech synthesis system have also demonstrated its robustness to non-ideal speech data that are recorded under varying conditions and with varying microphones, that are not perfectly clean, and/or that lack phonetic balance [2]. In fact we have demonstrated that we can create 1000s of TTS voices from non-TTS corpora such as ASR corpora and that can easily increases variability of speaker characteristics [3, 4]. This technique can produce applications that are beneficial in various domains. For example, it has a direct application in voice banking or voice reconstruction for patients who have or are threatened by throat cancer, or in the creation of alternative communication aids for patients with e.g. Parkinson's disease, in which the patient's original voice characteristics are preserved [5].

The 1000s TTS voices are available from an interactive online TTS demonstration system with a geographical representation which we devised recently[1]. The voices in this demonstration were built using pre-defined training recipes for each corpus. More importantly this device gave us good opportunities to compare the quality of synthetic speech for many speakers at the same time.

Careful listening revealed 1) that the quality of synthetic speech varies according to which corpus is used to train the average voice models, or by the amount of adaptation data used and 2) that there are a few speakers whose synthetic speech sounds worse than that of other speakers who have the same amount of adaptation data from within the same corpus.

For the first case, our previous analysis has already shown that the amount of adaptation data required for reproducing speaker similarity above a certain level varies by target speakers (and acoustic features) and ranges from three minutes to six minutes in terms of speech duration [6] and also that the naturalness of the synthetic speech generated from the adapted models is closely correlated with the amount of data used for training the average voice model [7]. We also know that gender-dependent average voice models provide better speaker adaptation performance than gender-independent average voice models for TTS [7]. This directly explains the relatively low quality of voices built on a small corpus (such as the RM corpus) since the small corpus does not satisfy the two conditions above.

The interesting phenomenon observed in the second case is new and analogous to the familiar situation in ASR, where WER varies widely across some speakers and is especially high for a small number of speakers [8]. In this paper we investigate this phenomenon from the point of view of TTS.

Initially we suspected the negative effects of recording condition mismatch since the acoustic differences due to inconsistent recording conditions were found to be greater than acoustic differences between speakers [3, 4]. During the analysis of the recording conditions/sites, however, we came across a new and meaningful finding for the phenomenon *by accident*, that is, a correlation between the naturalness of synthetic speech and the distance between the adapted speaker's model and the average voice model, instead of a correlation between recording conditions and naturalness of synthetic speech. Furthermore we remarked that this correlation is strongly linked to "vocal attractiveness."

2. HMM-based Speech Synthesis Systems and Experimental Conditions

The HMM-based speech synthesis system consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

In the speech analysis part, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [9]) mel-cepstral vocoder with mixed excitation (i.e., the mel-cepstrum, $\log F_0$ and a set of band-limited aperiodicity measures) are extracted as feature vectors for HMMs. In the average voice training part, context-dependent multi-stream left-to-right tied-state multi-space distribution hidden semi-Markov models are

[1]http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map.html

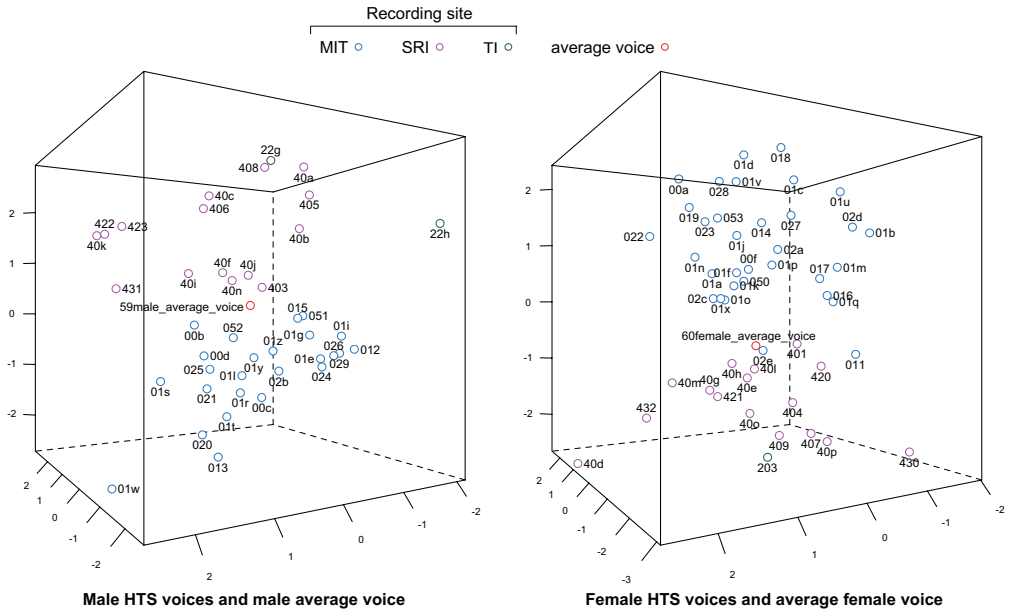

Figure 1: Multidimensional scaling of 120 HTS voices trained on the WSJ0 corpus. The three characters at each point correspond to the name of each speaker in the database. Left part shows the the male speakers and male average voice and right parts shows the female speakers and female average voice.

trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. A set of model parameters (mean vectors and diagonal covariance matrices of Gaussian pdfs) for the speaker-independent MSD-HSMMs is estimated using the EM algorithm. All EM re-estimation processes utilize speaker-adaptive training based on constrained maximum likelihood linear regression [10].

In the speaker adaptation part, the speaker-independent MSD-HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression [7]. In the speech generation part acoustic feature parameters are generated from the adapted MSD-HSMMs using a parameter generation algorithm that considers both the global variance of the trajectory to be generated and trajectory likelihood [11]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add. This signal is used to excite a mel-logarithmic spectrum approximation filter corresponding to the STRAIGHT mel-cepstral coefficients to generate the speech waveform.

Using the framework above, we built gender-dependent average voice models from short term, long term (excluding the speakers from very long term), development, and evaluation subsets of the WSJ0 corpus [12]. The numbers of training sentences are 10847 and 12151 sentences for male and female average voice models, respectively. They have 21.1 hours and 24.6 hours of speech respectively.

3. Visualization of 120 voices adapted and average voices built on the WSJ0 corpus using multidimensional scaling

We can place the voices created in this way in a low dimensional space derived from the properties of the speech which they generate and can visually analyze the distribution of speakers. There are several conventional approaches for visualizing speakers or speaking styles based on acoustic models or acoustic features [13, 14]. A similar visualization can be straightforwardly achieved using the HTS voices built and multidimensional scaling (MDS) [15].

Although we have already shown parts of this result in [3], the lower dimensional space is very important in the analysis of listening tests presented later and thus we reproduce the visualisation results here using more voices and the three-dimensional space.

Using all test sentences from the Blizzard Challenge 2008, we generated a set of speech samples from the gender-dependent average voice models and 120 HTS voices, each of which had a hundred adaptation sentences. We then calculated the average mel-cepstral distance between the speech for all pairs of voices, placing the values in mel-cepstral distance tables. For simplicity, the unadapted duration models of the average voice model were used so that the number of frames of synthetic speech for each speaker is the same. Then we applied a classic multidimensional scaling technique to the mel-cepstral distance table and examined the resulting three-dimensional space, which is shown in Figure 1. On the left-hand side of the figure, the MDS of the male speakers and male average voice appear and on the right, that of the female speakers and female average voice.

The axes of this space do not have any *pre-defined* meaning, but MDS attempts to preserve the pairwise distances between speakers given in the mel-cepstral distance table. In other words, similar speakers will be close to one another in this space. On examining the figure in detail, we noticed that all three-characters codes (corresponding to the names of speakers) distributed in the bottom part start with 0 and the codes for speakers distributed in top part start with 4. The first character of the names represents recording site for these speakers (0: MIT, 4:SRI, and 2:TI) [12]. Therefore we assigned different colors to each recording site in the figure.

It is apparent that recording conditions were not consistent among the recording sites although the same microphones were utilised. Furthermore, acoustic differences due to the inconsistent recording conditions are greater than acoustic differ-

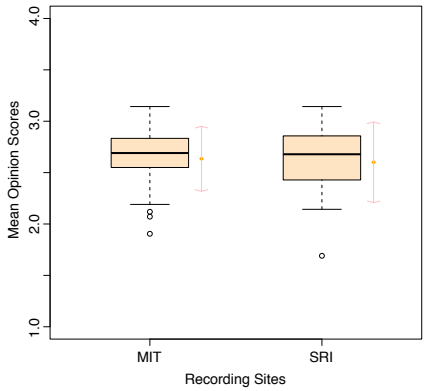

Figure 2: Standard box-plots are presented for evaluation scores of each site where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. Bar charts are presented for the word error rate interval data. In addition mean scores and their standard deviation are shown using arrows next to the box-plots.

ences between speakers since there is an obvious border between them.

4. Subjective evaluations of 59 voices adapted and an average voice

A natural next step for us is therefore to perform listening tests and to evaluate whether the acoustic differences due to the inconsistent recording conditions may cause fluctuation of the quality of synthetic speech generated from models adapted from the same average voice models using the same amounts of adaptation data.

For this purpose we utilized the same adapted voices and the same average voice used for MDS in the previous section and evaluated their naturalness using the MOS test in which four test sentences were randomly chosen from all the test sentences used for MDS above. The number of listeners was 40.

The score distributions for each site are shown in Figure 2, in which we cannot see clear differences between the results for each site. In fact, the Pearson product-moment correlation coefficient between the mean MOS scores obtained in the evaluation and the first axis of MDS which represents the recording sites is just -0.13. In a word, the MOS scores obtained are not correlated with the recording sites and associated recording condition differences. Interestingly the second axis of the MDS figure had somewhat stronger correlation (-0.38) than the first axis.

Therefore we decided we should examine other possible distances and focus on mel-cepstral distance between average voice and each voice, which can be viewed as a transformed distance of the voice. This correlation was stronger and it was -0.48. The fluctuation of the quality of synthetic speech was somewhat correlated inversely with mel-cepstral distance from the average voice. Its 95% confidence intervals are from -0.20 to -0.68.

Figure 3 shows the scatter plot of the mean MOS scores for the voices and the mel-cepstral distance from the average voice. This also represents a linear regression function fitted and its 95% confidence and prediction intervals. We can see that as the mel-cepstral distance from the average voice becomes larger, the MOS scores generally become worse. Readers might also be surprised that the average voice scores highest in the evaluation

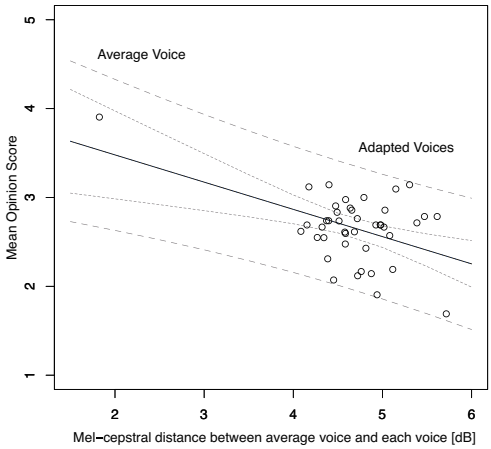

Figure 3: The scatter plot of the mean MOS scores of 59 male voices adapted and a male average voice model. Each dot represents either the male speaker or male average voice. Horizontal axis shows the mel-cepstral distance from the average voice. Vertical axis shows the mean MOS score obtained for each voice. This also represents a linear regression function fitted and its 95% confidence and prediction intervals. For computation of the mel-cepstral distance for the average voice itself, random-sampling-based parameter generation algorithm [16] was used.

(the mean MOS score is 3.9.). A similar trade-off phenomenon between transformed distance and quality reduction of synthetic speech has been observed even in voice conversion [17].

The correlation obtained is not strong enough. This explains only 23% of the behavior of the adapted voices and 77% is still unknown. However this becomes an important factor for determining how to train average voice models from many speakers. For instance, this could explain why gender-dependent average voice models provide better speaker adaptation performance than either gender-independent average voice models or speaker-dependent models for TTS. In addition, for achieving a better quality of synthetic speech based on our analysis results, this also implies that we may use multiple gender-dependent average voice models and may choose the nearest model if a huge amount of data is available. We note that all of them must have a sufficient quantity of training data since the amount of data for the average voice models is the most dominant factor for the quality of synthetic speech.

5. Average voice sounds more attractive than individuals?

In addition to the transformed distance mentioned in previous section, we hypothesize that there is a psychological reason.

It is well known that Langlois and Roggman have shown that averaged faces look more attractive than individuals in their paper entitled “Attractive Faces are Only Average” [18]. In a similar way, a likely psychological explanation for the higher score of the average voices is that *attractive voices are also average*. This is a very interesting aspect which has a deeper meaning and implies a new direction for the statistical parametric approach to speech synthesis since the statistical averaging effect, which is an acknowledged weakness of current HMM-based speech synthesisers, might have the potential to produce voices that sound more attractive than individuals.

A very recent psychoacoustic study [19] by Belin’s group

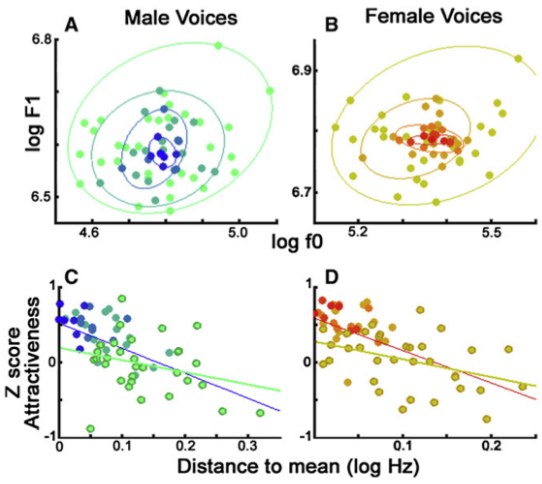

Figure 4: "In the logf0-logF1 space, Euclidean distance to mean was negatively correlated to vocal attractiveness rating (r=-0.59, adjusted R^2=0.34, $p<0.001$)." This figure is taken from [19].

verified the hypothesis using many speakers' vowels and their averaged vowels. Surprisingly they also found that their listening test scores are correlated with distance to the average voices as shown in Figure 4, whereas there are some differences between their experiments and our experiments:

- They used vowels only whereas we used sentences.
- We had only two average voices whereas they evaluated various combinations of speakers for constructing several average voices.
- They adopted Z scores on attractiveness rather than MOS on naturalness.
- Log F_0/F_1 space was used instead of mel-cepstral space.
- Large gap between average voices and adapted voices in our experiments. This may be explained by the recording condition inconsistency of our data. Our average voice models are located at the center of recording conditions rather than the center of the speakers due to the inconsistent recording conditions as can be seen from Fig. 1.

From the similarity of the tendency, we need to consider if there is a possibility that our listeners took vocal naturalness and attractiveness together. It leaves no doubt, however, that the averaging across multiple speakers has a positive effect on the speech produced by the statistical parametric approach to speech synthesis.

6. Conclusions

In speaker-adaptive HMM-based speech synthesis, there are a few speakers whose synthetic speech sounds worse than that of other speakers who have the same amount of adaptation data from within the same corpus. This paper has investigated this fluctuation in quality and has found that as mel-cepstral distance from the average voice becomes larger, the MOS scores generally become worse. Although the negative correlation obtained is not strong enough, this helps us improve the training and adaptation strategies of the average voice models. Furthermore we remark that this correlation is strongly linked to "vocal attractiveness." We believe this suggests an interesting new direction for statistical parametric speech synthesis.

Acknowledgements The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org).

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[3] J. Yamagishi *et al.*, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech 2009*, Brighton, U.K., Sep. 2009, pp. 420–423.

[4] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Speech, Audio & Language Process.*, 2010, (in press).

[5] S. Creer, P. Green, S. Cunningham, and J. Yamagishi, "Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit," in *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, J. W. Mullennix and S. E. Stern, Eds. IGI Global, Jan. 2010.

[6] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[7] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[8] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, "1993 benchmark tests for the ARPA spoken language program," in *HLT '94: Proceedings of the workshop on Human Language Technology*. Morristown, NJ, USA: Association for Computational Linguistics, 1994, pp. 49–74.

[9] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[11] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[12] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, New York, 1992, pp. 357–362.

[13] M. Shozakai and G. Nagino, "Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models," in *Proc. ICSLP 2004*, Jeju Island, Korea, Oct. 2004, pp. 717–720.

[14] A. Maier, M. Schuster, U. Eysholdt, T. Haderlein, T. Cincarek, S. Steidl, A. Batliner, S. Wenhardt, and E. Nöth, "QMOS – a robust visualization method for speaker dependencies with different microphones," *Journal of Pattern Recognition Research*, vol. 1, pp. 32 – 51, 2009.

[15] T. Cox and M. Cox, *Multidimensional Scaling*. Chapman and Hall, 2001.

[16] K. Tokuda, H. Zen, and T. Kitamura, "Reformulating the HMM as a trajectory model," *IEICE technical report. Natural language understanding and models of communication*, vol. 104, no. 538, pp. 43–48, Dec. 2004.

[17] D. Erro, "Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models," Ph.D. dissertation, Universitat Politecnica de Catalunya, 2008.

[18] J. H. Langlois and L. A. Roggman, "Attractive faces are only average," *Psychological Science*, vol. 1, no. 2, pp. 115–121, 1990.

[19] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G. A. Rousselet, H. Kawahara, and P. Belin, "Vocal attractiveness increases by averaging." *Current Biology*, vol. 20, no. 2, pp. 116–120, 2010.